Performance of the Federal Standard 2.4 kbps MELP Vocoder Over ATM Networks

Deepak Sridhara and Thomas Fuja [1]
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556
e-mail: {dsridhar, tfuja} @nd.edu

Abstract — **This paper evaluates the performance of the federal standard 2.4 kbps MELP vocoder when used over ATM networks that are subject to cell loss. It begins by addressing the sensitivity of MELP to random deletion of frames; it quantifies the extent to which the quality of the reconstructed speech is affected by two factors – the rate at which frames are lost and the burstiness of that loss process. Next, the ATM structure is introduced and techniques for ameliorating the effects of ATM cell loss when transporting MELP-compressed speech are evaluated. Interleaving MELP frames into ATM cells is shown to improve performance significantly. Finally, the effectiveness of erasure decoding using Reed-Solomon codes to recover lost cells is investigated. Throughout this paper, cell loss is simulated by an i.i.d. model and also a bursty (Gilbert) model. The quality of the reconstructed speech is assessed using both objective measures (i.e., frequency-weighted average spectral distortion and percentage of outliers) and subjective tests.**

I. Introduction

Packet network architectures such as asynchronous transfer mode (ATM) technology have been designed to transport multimedia signals, as well as data, at high rates. As voice may well constitute a significant portion of the traffic in these networks, reliable transmission of voice in such environments is of considerable importance. The conflicting goals of conserving bandwidth and achieving high rates of transmission suggest the use of compressed speech over these networks. This paper evaluates the performance of the federal standard 2.4 kbps mixed excitation linear predictive (MELP) vocoder for this purpose.

In an ATM system, packets (called *cells*) from different sources are statistically multiplexed at a switching node and subsequently transported over high speed links. When the rate of arrival of cells from different sources increases, the resulting "traffic congestion" may cause buffers at the switches to overflow. Under these circumstances, some of the cells may have to be discarded. For real-time speech applications, retransmission of lost cells containing speech frames is not feasible due to the delay involved. As an alternative, efficient estimation and reconstruction of such missing cells at the receiver is important.

[1]This work was supported by the U.S. Department of Defense under contract MDA904-99-C-2608.

The MELP vocoder described in [1] is based on the traditional linear predictive coding (LPC) model with a number of additional features; these features include a mixed excitation signal (i.e. a mixture of noise and pulse excitation used as an input to the synthesis filter), an adaptive spectral enhancement filter, Fourier magnitude modeling of the pulse excitation, and a pulse dispersion filter. The MELP vocoder encodes 22.5 msec of speech to a 54-bit frame. Most of the parameters in a typical frame are quantized using suitable vector codebooks, and only the codebook indices are transmitted.

An ATM cell can be viewed as a 53 byte fixed length packet that contains a five byte ATM header and a four byte adaptation layer header [6][8]; this allows a maximum of six 54-bit MELP-encoded frames to be packed into every ATM cell. When the ATM cells are subject to loss due to traffic congestion, this causes a concurrent loss of the MELP frames; the nature and the effect of the MELP frame loss process depends on how the MELP frames are "packed" into the ATM cells.

In this paper, the following issues are addressed:

- The robustness of the MELP decoder to random deletion of MELP frames. Two characteristics of the loss process - the loss-rate and the burstiness of the loss - are identified as the most significant factors affecting the quality of the synthesized speech. Techniques such as frame-repetition and frame-interpolation for reconstructing the missing frames are implemented and compared.
- The sensitivity of MELP-encoded frames that are loaded into ATM cells to random cell losses. The improvement in performance due to interleaving frames in cells is shown. This improvement is achieved as a consequence of reducing burstiness among missing frames; an analysis of the burstiness generated in relation to the cell loss-rate and the interleaving-depth is presented.
- Finally, the use of Reed-Solomon codes to ameliorate cell loss. When the cell loss rate is sufficiently high, the use of Reed-Solomon codes and "erasure correction" may be necessary to provide acceptable speech quality. Codes of different rates are used in conjunction with the ATM format to give substantial improvements in performance. The tradeoffs between the gain in performance and the efficiency of the code are highlighted.

The rest of this paper is organized as follows. The sensitivity of the MELP vocoder to frame deletion is assessed in Section II. Section III discusses the performance of MELP in ATM networks; the effectiveness of interleaving to reduce the burstiness of frame losses in ATM systems is investigated here. Finally, in Section IV the use of Reed Solomon codes

to improve cell loss in ATM is presented. Conclusions and possible future work are described in Section V.

II. Sensitivity of MELP to frame deletion

It is important to understand the sensitivity of MELP subject to frame loss as a basis for evaluating its performance over networks that are subject to cell loss. The loss of a highly compressed speech frame is substantial when compared to the loss of an uncompressed frame; the duration of speech lost is greater in the former case. MELP, being a low rate vocoder, is very likely to be highly sensitive to frame loss. This is a motivation for finding appropriate schemes that will reduce the effects of distortion when MELP frames are subject to deletion.

When a MELP frame is lost and the receiver must deal with that loss, there are (at least) three obvious approaches:

1. frame-muting, i.e. the lost speech is merely replaced by silence;
2. frame-repetition, whereby the missing frames are replaced by the frame immediately before the missing frames; and
3. frame-interpolation, whereby the missing frames are interpolated using the frames that are immediately prior to and after the missing frames.

When the missing frames are merely muted, there is no effort made towards reconstruction; the quality of the resulting speech is therefore a direct measure of the sensitivity of this vocoder to frame loss and also serves as a yardstick against which the improvements offered by other reconstruction methods can be measured. When the transitions in speech across the missing frames are small, frame-repetition suffices in replacing the lost speech with a close sounding signal; this is especially true when the burstiness of the loss (i.e. the number of missing frames that need to be replaced) is small. Frame-interpolation involves interpolating the speech parameters of the end frames across the deleted frames. Interpolation has an effect of inducing smooth signal transitions across the missing frames; here again, the shorter the burst, the better is the smoothing effect. The frame loss is concealed to a fair degree when either interpolation or repetition have to be performed over an isolated frame loss.

Intuitively, frame loss can be characterized by two properties - the fraction (or "rate") at which frames are lost and the "burstiness" of the loss generated. When the loss-rate is low, reconstruction is relatively rare and may pass unnoticed. When the loss exhibits low burstiness, then the losses tend to be fairly isolated and so either repetition or interpolation of parameters may be adequate to avoid unacceptable distortion. It is therefore important to relate the quality of the output obtained with the frame loss-rate and the burstiness of frame loss.

A 3:18 min speech segment containing 8800 MELP frames was used to assess the sensitivity of MELP to frame loss[2]. Approximately half of the MELP frames contained speech by male speakers and half by female speakers. The encoded frames were then subject to random deletion; each frame was deleted with independent probability p. Quality of speech was measured for the output obtained when the missing speech was reconstructed using the techniques described above.

[2]The same speech segment was used in every simulation described in this paper.

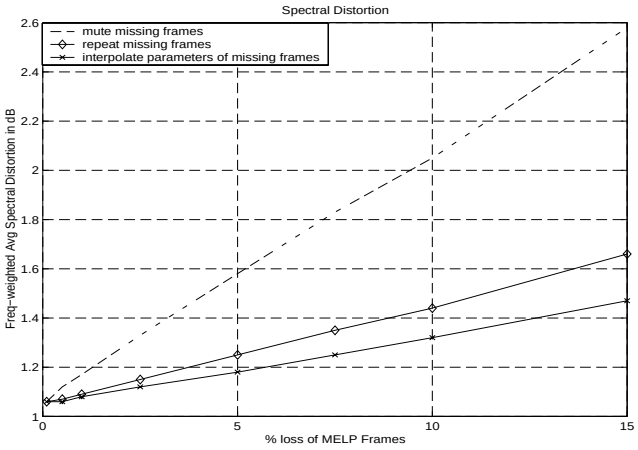
Spectral Distortion
mute missing frames
repeat missing frames
interpolate parameters of missing frames
Freq−weighted Avg Spectral Distortion in dB
% loss of MELP Frames

Figure 1: Frequency-weighted average spectral distortion assuming each MELP frame is dropped with independent probability p

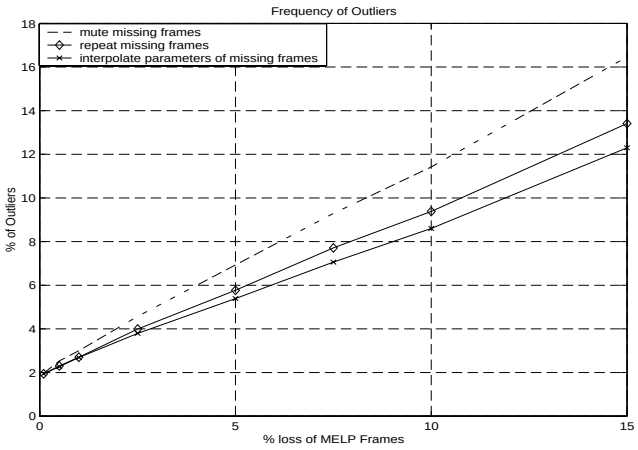

Figure 2: Fraction of outliers - fraction of frames whose spectral distortion exceeds 2.0 dB, assuming each MELP frame is dropped with independent probability p

As was conjectured, the vocoder was found to be highly sensitive to frame loss. The distortions computed by objective measures [2][10] specified for this vocoder are shown in Figures 1 and 2. They indicate that the output deteriorates almost in direct proportion to the loss-rate. It's clear that parameter repetition and interpolation both produce better speech than muting; this is clear from the graphs in terms of objective measures and it is also observed in listening tests. Moreover, parameter interpolation slightly outperforms parameter repetition.

There is no delay involved in filling the deleted frames by frame-repetition, but frame-interpolation incurs a delay of one frame in the best case; this is the processing time required for decoding the end frames needed for interpolation.

It was noted earlier that the *burstiness* of the loss process – in addition to the *rate* at which frames are lost – may play a significant role in the decoder's ability to reconstruct the missing frames. Here, burstiness refers to the tendency of frame losses to be grouped together in time – i.e., as they would be during times of traffic congestion at a switch. To isolate the effect of burstiness from the effect of frame loss rate, it is nec-

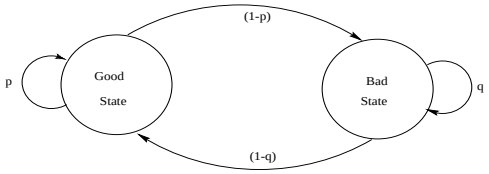

Figure 3: Gilbert model to simulate burst of cell/frame discards

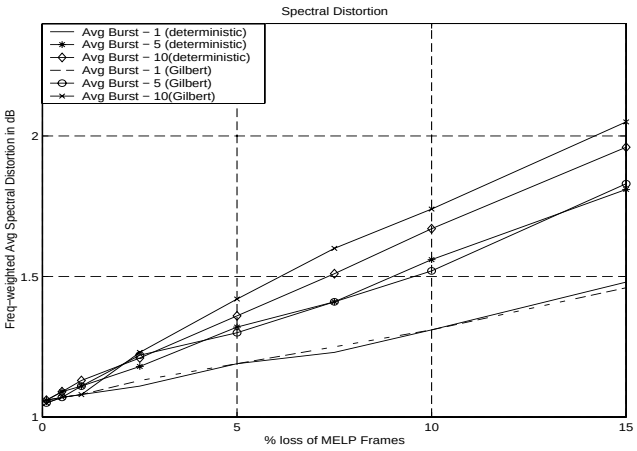
Spectral Distortion
Avg Burst – 1 (deterministic)
Avg Burst – 5 (deterministic)
Avg Burst – 10(deterministic)
Avg Burst – 1 (Gilbert)
Avg Burst – 5 (Gilbert)
Avg Burst – 10(Gilbert)
Freq–weighted Avg Spectral Distortion in dB
% loss of MELP Frames

Figure 4: Frequency-weighted average spectral distortion for the two bursty models - one deterministic and the other probabilistic

essary to develop models for frame loss that permit different levels of burstiness for the same loss rate. The following two models were used:

- A deterministic model, in which the first L frames are dropped and the next $L(1-p)/p$ frames retained - this process being repeated over and over. This model simulates bursts of length L with an effective frame drop rate of p.
- A probabilistic Gilbert model [9] in which there are two possible channel states associated with each MELP frame - a *good* state and a *bad* state (see Figure 3); a frame is never deleted in the good state, but always deleted in the bad state. Under this model, a transition occurs from the good state back to the good state with probability p (and from the good state to the bad state with probability $1-p$); similarly, a transition from the bad state back to the bad state occurs with probability q. The average burstlength with this model is $1/(1-q)$, and the loss rate for the Gilbert model is $(1-p)/(2-p-q)$.

The above channel models were simulated and frame-interpolation was used for reconstruction of deleted frames. The quality of the reconstructed speech is found to deteriorate significantly at higher burstlengths when the frame loss-rate is held constant. An isolated loss (a "burst" of length one) is almost always well-concealed by interpolation, especially at loss rates below ten percent. Objective measures in Figures 4 and 5 indicate the effects of burstiness on the synthesized speech. The two models yield similar performance, indicating that the relevant issue is indeed burstlength and not tied to the particular model.

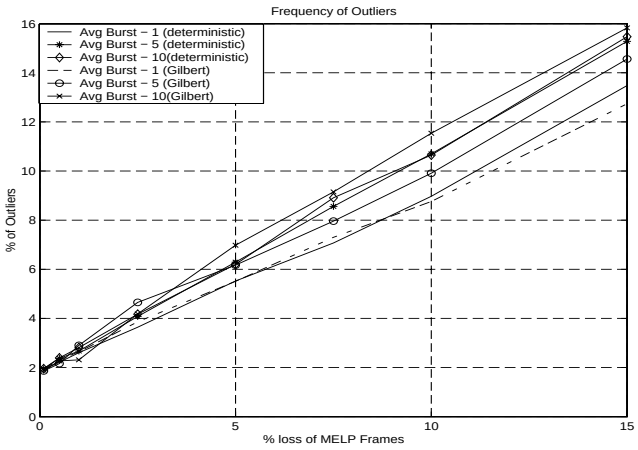
Frequency of Outliers
Avg Burst – 1 (deterministic)
Avg Burst – 5 (deterministic)
Avg Burst – 10(deterministic)
Avg Burst – 1 (Gilbert)
Avg Burst – 5 (Gilbert)
Avg Burst – 10(Gilbert)
% of Outliers
% loss of MELP Frames

Figure 5: Fraction of Outliers for the two bursty models - one deterministic and the other probabilistic

These results show that the distorting effects of loss can be mitigated by limiting the burstiness of loss even at fairly high loss rates. Thus it is clear that loss rate and loss burstiness are two of the most important characteristics of the frame loss process. This lays the groundwork for combating cell loss (and thereby, frame loss) in ATM networks.

III. MELP Over ATM

Performance of MELP in ATM framework is now considered. A maximum of six frames can be packed into a single ATM cell; when six consecutive MELP frames are loaded into every ATM cell, the loss of a single cell will cause a burst of six consecutive MELP frames. Since the frame loss *rate* is assumed (at least initially) to be beyond our control – the frame loss rate will be equal to the cell loss rate when equal number of frames are packed into every cell and no error control codes are used – our only hope in controlling the effect of cell loss is to reduce the burstiness of the frame loss process. The obvious approach is to interleave MELP frames within the cells.

Figure 6 indicates how frames can be block interleaved into the ATM structure at different interleaving depths. When no interleaving is used, loss of a single isolated cell results in the loss of six consecutive frames; when an interleaving depth of N is used, loss of an isolated cell results in the loss of six frames that are each separated by $(N-1)$ available frames.

The usefulness of interleaving can be measured by randomly deleting ATM cells containing MELP frames and assessing the performance of interleaved versus non-interleaved processing. (Interpolation of parameters was used to replace missing MELP frames.) Substantial gains in the speech quality are obtained when frames are interleaved, as seen from Figures 7 and 8. Note that substantial gain is obtained when interleaving to depth two is employed compared with the no-interleaving case; however, the increase in performance is marginal when the interleaving depth is increased from two to four. Intuitively, this only implies that, at an i.i.d. cell loss rate of 10-15 percent, an interleaving depth of two is sufficient to isolate most of the missing frames. The same may not be true for a more bursty cell-loss model.

The most common measure of burstiness is the average burstlength – i.e., how many *consecutive* MELP frames are lost every time *at least one* MELP frame is lost, on the average.

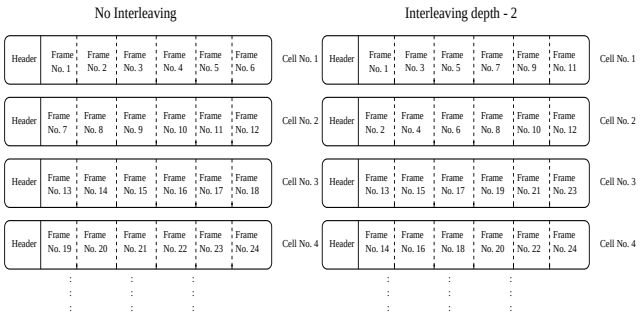

Figure 6: Cells containing frames that are not interleaved and that are interleaved to depth two

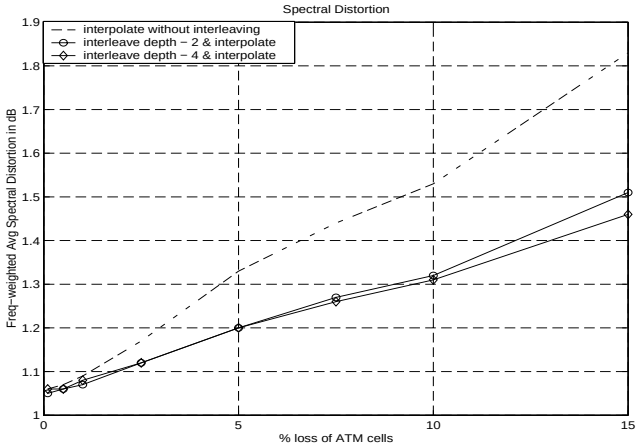

Figure 7: Frequency-weighted average spectral distortion when ATM cells are dropped with independent probability p

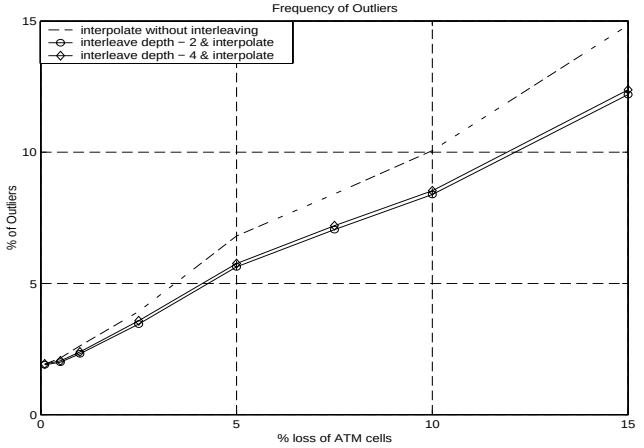

Figure 8: Fraction of Outliers when ATM cells are dropped with independent probability p

It is possible to derive this average burstlength assuming an i.i.d. loss of ATM cells. Two particular cases are considered - one without interleaving and the other in which MELP frames are interleaved to depth two into the ATM cells.

When no interleaving is used, the analysis is simple. The only place that bursts can start are in the first MELP frame in each ATM cell - i.e., MELP frames numbered $1, 7, 13, \ldots$. Furthermore, every burst will last for a multiple of six MELP frames. If the ATM cells are lost with independent probability p, then the probability that a burst will last $6m$ frames is just $p^m(1-p)$, meaning that the expected burstlength is

$$\overline{B} = \frac{6}{1-p}.$$

To analyze the system with interleaving, it is useful to define the random process $\{X_i : i = 1, 2, \ldots\}$, where

$$X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ MELP frame is lost} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Furthermore, define the random process $\{B_i\}$ in such a way that B_i indicates the length of a burst starting with the i^{th} frame - i.e.,

$$B_i = \begin{cases} 0 & \text{if } X_{i-1} = 1 \text{ or } X_i = 0; \\ k & \text{if } X_{i-1} = 0,\ X_i = X_{i+1} = \ldots = X_{i+k-1} = 1, \\ & \quad \text{and } X_{i+k} = 0. \end{cases} \tag{2}$$

So B_i is non-zero if and only if a burst begins with the i^{th} MELP frame. The expected burstlength may be derived from the conditional distribution $P(B_i = k | B_i > 0)$ for $k = 1, 2, \ldots$. This distribution depends on i; more specifically, it depends on i mod 12. This conditional distribution is written more specifically as,

$$\begin{aligned} P(B_i = k | B_i > 0) &= P(X_{i-1} = 0, X_i = X_{i+1} = \ldots \\ &\quad = X_{i+k-1} = 1, X_{i+k} = 0 \,\big|\, X_{i-1} = 0, X_i = 1) \end{aligned} \tag{3}$$

These probabilities are easily computed when cells are deleted with independent probability p. For example, $P(B_i = 1 | B_i > 0) = 1 - p$ when $i \equiv 1 \bmod 12$.

$E[B_i | B_i > 0]$ is then found to be

$$E[B_i | B_i > 0] = \begin{cases} \frac{1+p+10p^2}{1-p^2} & \text{if } i \equiv 0 \bmod 12 \\ \frac{1+11p}{1-p^2} & \text{if } i \equiv 1 \bmod 12 \\ 1 & \text{if } i \equiv (2, 3, \ldots \text{ or } 11) \bmod 12 \end{cases} \tag{4}$$

If we now average over the twelve equivalence classes of i (modulo 12), we arrive at the average burstlength for the case where depth-two interleaving is used:

$$\overline{B} = \frac{1}{1-p}.$$

This analytical result matches the simulation results. Similar analysis could extend the results to larger interleaving depths as well as different cell loss models.

IV. Use of Reed-Solomon Codes to Recover Discarded Cells

The results obtained up to this point have shown that interleaving is effective in controlling the burstiness of MELP frame loss in an ATM environment, and this reduction of burstiness improves the quality of the recovered speech. However, the useful effects of interleaving are limited at high cell loss rates, and this suggests the use of redundancy - i.e., error control codes - to recover lost cells. In particular, the use of Reed-Solomon codes in the ATM format [5] to provide for erasure-correction of missing cells is considered.

An (n, k) Reed-Solomon (RS) code over GF(q) takes k data symbols from a q-ary alphabet and adds $n-k$ symbols of redundancy; as a result, every pair of valid n-symbol codewords differs in at least $d_{\min} = n-k+1$ places - i.e., the code has a *minimum distance* of $n-k+1$. Such a code is capable of either correcting up to $\lfloor (d_{\min}-1)/2 \rfloor$ symbol errors or recovering up to $d_{\min}-1$ symbols that have been deleted ("erased") from an n-symbol codeword. As this paper doesn't consider the effects of noise in the channel, the cells that aren't lost due to traffic congestion are assumed to be error-free. Therefore, the appropriate use of Reed-Solomon codes in this context is "pure" erasure correction.

Figure 9 shows how RS encoding may be performed across ATM cells. Every cell is loaded with six consecutive MELP frames. Different (n, k) Reed-Solomon codes over GF(256) are used. The first byte from the data payloads of cells 1 through k are encoded to generate $(n-k)$ parity symbols (or bytes); the next $n-k$ cells are marked as redundant, with the $n-k$ parity bytes generated from the above encoding loaded into these cells. The same encoding is repeated for the rest of the data bytes in the cells. This is then repeated for every set of k original cells containing MELP frames; as a result, the number of ATM cells needed for carrying encoded speech is increased by a factor of n/k.

Simulations using $(32, k)$ codes were performed for different values of k. The missing cells are marked as "erasures" and erasure-correction is performed. The cells that were not recovered from erasure-decoding are reconstructed by parameter-interpolation across deleted frames. The performance of different rate RS codes is evaluated. A lower rate code is certainly capable of recovering a greater number of deleted cells; the obvious tradeoff is the efficiency of the code, i.e. more overhead is required to transmit the same information.

Figures 10 and 11 show the objective performance of the test sequence assuming an i.i.d. model and Reed-Solomon codes of varying rate. It is seen that, with a code rate of $20/32 = 0.625$ (60 % overhead) it is possible to achieve near-noiseless performance even when up to 15% of the ATM cells are dropped.

Section II quantified the extent to which the receiver's ability to reconstruct MELP-encoded speech is dependent upon the rate at which MELP frames are lost. An issue of obvious interest is the "effective rate" at which the receiver "sees" frames being lost when Reed-Solomon codes are used over a channel where ATM cells are being lost at a given rate.

Suppose n ATM cells are transmitted over a network - k MELP-bearing cells and $n-k$ redundant cells. Suppose further that each ATM cell is lost with independent probability p. Then the probability that exactly i cells (of the n) are lost and that, further, of those i cells, a of them are MELP-bearing is given by

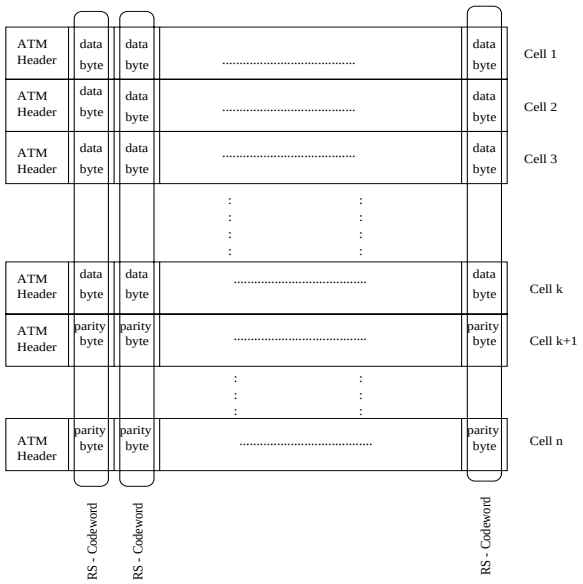

Figure 9: Reed Solomon codes in ATM cells

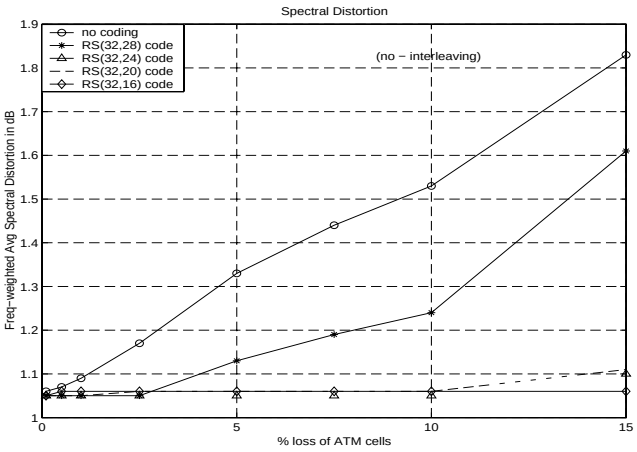

Figure 10: Frequency-weighted average spectral distortion when ATM cells are dropped with independent probability p

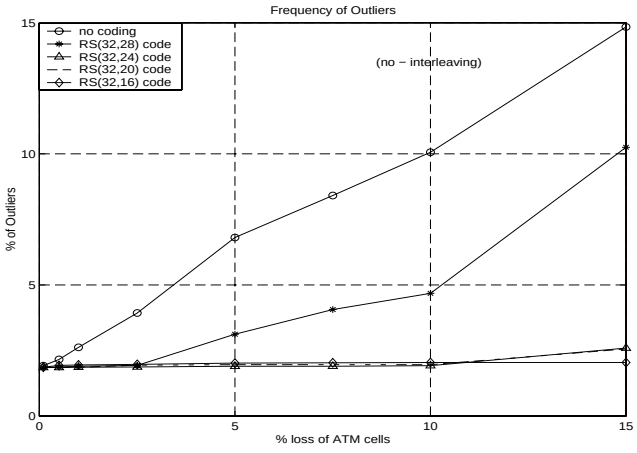

Figure 11: Fraction of outliers when ATM cells are dropped with independent probability p

$$P_i(a) = \binom{k}{a}\binom{n-k}{i-a}p^i(1-p)^{n-i} \quad (5)$$

The average number of MELP cells that are not recovered after erasure-correction is therefore

$$E[A] = \sum_{i=n-k+1}^{n} \sum_{a=1}^{min(i,k)} a\binom{k}{a}\binom{n-k}{i-a}p^i(1-p)^{n-i} \quad (6)$$

The rate at which MELP frames are lost is thus

$$P_{eff} = \frac{1}{k}\sum_{i=n-k+1}^{n} \sum_{a=1}^{min(i,k)} a\binom{k}{a}\binom{n-k}{i-a}p^i(1-p)^{n-i} \quad (7)$$

Similar analysis can be performed for cell losses generated by the Gilbert model.

V. Conclusions and Future work

In this paper, it has been shown that the rate at which MELP frames are lost and the burstiness of the loss process are the two most significant factors affecting a MELP decoder's ability to reconstruct the original speech. Interpolation of MELP parameters across the missing frames is found to mitigate the effects of loss. The burstiness of frame loss may be controlled by interleaving frames within ATM cells. Interleaving is clearly beneficial at low cell loss rates. Using Reed Solomon codes at the cost of some ATM overhead is found profitable at higher loss rates.

Not included in this paper but of interest is the use of ATM cell prioritization to guarantee that, when ATM cells are lost, they will contain MELP parameters that affect speech quality to the least degree possible. Another topic of interest is the use of *residual redundancy* in the MELP bitstream [7] to aid in the interpolation of missing MELP frames.

Acknowledgments

The authors would like to thank Doug Rahikka for stimulating their interest in this subject and for numerous interesting conversations. They would also like to thank the group maintaining the speech corpus at Linguistic Data Consortium, University of Pennsylvania, for access to their database for the simulations in this paper.

References

[1] DoD Draft - "Specifications for the Analog to Digital Conversion of Voice by 2,400 Bit/Second Mixed Excitation Linear Prediction" May 28,1998.

[2] A. McCree, K. Truong, E. B. George, T. Barnwell, and V. Viswanathan, "A 2.4 kbit/s MELP Coder Candidate for the New U.S. Fedaral Standard," *Proc. of the 1996 Intl. Conference on Acoustics, Speech, and Signal Processing*, pp. 200-203.

[3] Kang and Franmsen, "MELP Error Protection by ANDVT HF Modem", *Voice Systems Section, Naval Research Laboratory* , August 1996.

[4] S. Lin and D. Costello, *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, 1993.

[5] Y. Rasheed and A. Leon-Garcia, "AAL1 with FEC for the Transport of CBR MPEG2 Video over ATM Networks", *Proceedings of the 1996 IEEE INFOCOM.*

[6] Kotikalapudi Sriram, R. Scott McKinney, Mostafa Hashem Sherif, "Voice Packetization and Compression in Broadband ATM Networks", *IEEE Journal on Selected Areas in Communications*, April 1991, pp. 294-304.

[7] T. Fazel, "Design of Joint Source-Channel Decoding Algorithm for the 2400 bps MELP Speech Coder", *A PhD. Dissertation, UMCP, Maryland*, Dec 1999.

[8] David J. Wright, "Voice over ATM: An Evaluation of Implementation Alternatives", *IEEE Communications Magazine*, May 1996.

[9] Hiroshi Ohta and Tokuhiro Kitami, "Simulation Study of the Cell Discard Process and the Effect of Cell Loss Compensation in ATM Networks", *The Trans. of the IEICE*, Vol. E 73, No. 10, Oct 1990.

[10] D. Sridhara, "Performance of Federal Standard 2.4 kbps Speech Coder over ATM networks" - Master's dissertation in preparation.